LLNL-JRNL-540291

Limit of detection of Bacillus anthracis in complex soil and air samples using next-generation sequencing

N. Be, J. Thissen, S. Gardner, K. McLoughlin, V. Fofanov, H. Koshinsky, T. Brettin, P. Jackson, C. Jaing

March 21, 2012

Public Library of Science (Plos)

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Limit of detection of *Bacillus anthracis* in complex soil and air samples using next-generation sequencing

Nicholas A. Be[1], James B. Thissen[1], Shea Gardner[1], Kevin McLoughlin[1], Viacheslav Fofanov[2], Heather Koshinsky[2],
Tom Brettin[3], Paul Jackson[1], and Crystal Jaing[1]

[1]Global Security, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA
[2]Eureka Genomics, Hercules, CA 94547, USA
[3]Oak Ridge National Laboratory, Oak Ridge, TN

Running title: Detection of *B. anthracis* by next-generation sequencing

Corresponding author: Crystal Jaing, Ph.D.
Biosciences and Biotechnology Division
Lawrence Livermore National Laboratory
Livermore, CA 94551
Tel: 925-424-6574
Email: jaing2@llnl.gov

LLNL-JRNL-540291

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

ABSTRACT

Bacillus anthracis is the potentially lethal etiologic agent of anthrax disease, and is a significant concern in the realm of biodefense. One of the cornerstones of an effective biodefense strategy is the ability to detect infectious agents with a high degree of sensitivity, particularly in the context of a complex sample background. The genomic structure of *B. anthracis*, however, renders specific detection difficult due to close homology with its less clinically destructive near neighbors, *B. cereus* and *B. thuringiensis*. We therefore elected to determine the efficacy of next-generation sequencing for the detection of *B. anthracis* within an environmental background. We employed the Illumina and 454 sequencing platforms to sequence titrated genome copy numbers of *B. anthracis* in the presence of background nucleic acids from both aerosol and soil material. We found high-throughput sequencing to be capable of detecting as few as 10 genome copies per ng background nucleic acid. Detection was effectively accomplished both by mapping reads to a defined subset of reference genomes and by comparison to the full GenBank database. Further, by assessing multiple bioinformatics analysis methods, we found that sequencing data obtained from *B. anthracis* could be reliably distinguished from both *B. cereus* and *B. thuringiensis*. Finally, we also tested and compared the efficacy of a microbial census microarray for detection of *B. anthracis*. Our results demonstrate that next-generation sequencing provides a sensitive and specific method for the analysis of *B. anthracis* in a complex sample background. This, in combination with the comprehensive ability of sequencing to identify novel and unique genetic variants, indicate that such protocols should represent an important component of any biosurveillance strategy.

Keywords: *Bacillus anthracis*, anthrax detection, next-generation sequencing, detection microarray biodefense, biosurveillance

INTRODUCTION

Bacillus anthracis is the gram-positive etiologic agent of the potentially fatal infectious disease anthrax. *B. anthracis* exhibit the ability to form dormant endospores resistant to extreme environmental conditions and can persist for long periods of time in terrestrial or aquatic environments. Spores function as infectious agents by exiting the dormant state upon contact with a nutrient-rich environment, which, in the case of humans, occurs following respiratory or cutaneous exposure. This germination process subsequently leads to the expression of toxins responsible for virulence in humans and other mammals [1]. These toxins, along with factors required for bacterial encapsulation, are encoded on the plasmids pXO1 and pXO2 [2]. Two of the anthrax toxin proteins, edema factor (EF) and lethal factor (LF), form complexes with protective antigen (PA), to cause severe edema and cell death when merged in binary combinations [3].

The acute and potentially lethal nature of infection with *B. anthracis*, in combination with the resilience of its vegetative spores upon exposure to heat and radiation, contributes significantly toward the possibility of its use as an aerosolized bioweapon. During the months of September to November of 2001, *B. anthracis* spores were sent via public mail to government and news organizations, resulting in 22 cases of anthrax and 5 deaths [4]. Members of the public health field have, in the decade following these attacks, exhibited amplified interest and investment in surveillance for agents of bioterrorism, particularly *B. anthracis*. These efforts require significant collaboration between microbiologists, clinicians, public health officials, and law enforcement.

One of the most important aspects of an effective response to a biologically-based attack is the ability to detect potentially infectious agents with a high degree of sensitivity and specificity. The LD_{50} for inhalation anthrax is estimated to be approximately 8,000 colony-forming units (CFU), although the minimal infectious dose is expected to be much lower [5]. Successful detection technologies will, therefore, need to be capable of identifying *B. anthracis* bacilli within this range. An additional major challenge to *B. anthracis* detection lies in its close relation to *B. thuringiensis* and *B. cereus*. The high degree of sequence similarity within this group and the possibility of horizontal plasmid transfer pose significant challenges when attempting to distinguish species within this genus.

A number of methods, ranging from basic to complex in their application, have been employed for the detection of *B. anthracis*. Most simply, the bacilli may be cultured and identified on blood agar. Microbiological techniques are, however, slow and require personnel trained in bacillary morphology, in addition to biosafety level 3 (BSL-3) facilities. Additionally, culture cannot provide genetic detail, with the exception of limited antibiotic resistance information. Identification via biochemical characteristics has been examined, including assays for lipids characteristic of certain bacilli [6] and intact cell mass spectrometry for the construction of a spectrum typifying *B. anthracis* [7]. Such methods, however, often require culture and depend on the quality of available lipid and proteomic databases [8]. Numerous immunoassays have been developed for the detection of *B. anthracis* antigens and toxins [9-11], but their utility is limited by low sensitivity and reduced specificity due to significant homology with *B. cereus* [11]. Each of the above techniques is also limited in potential for characterization of novel or unknown variants.

The increased capacity and reduced cost of next-generation sequencing have resulted in their increased application toward microbial identification and characterization. The ability to sequence full bacterial genomes provides the opportunity for heightened specificity when comparing near genetic neighbors. In order to assess the efficacy of deep sequencing for detection of *B. anthracis* in the environment, we added defined copy numbers of the *B. anthracis* genome to whole nucleic acid from both aerosol and soil particulates, simulating a complex environmental background. We sequenced this material using both Illumina and 454 platforms and processed the resultant data using multiple bioinformatics approaches. Further, we determined specificity by examining detection of the closely related *B. thuringiensis* and *B. cereus*. Finally, we compared our sequencing results to a previously developed microbial census microarray.

RESULTS

Detection of B. anthracis *in environmental samples by Illumina sequencing and mapping to reference subsets*

We inoculated *B. anthracis* genomic DNA into background DNA extracted from either aerosol filter particulates or soil samples and performed whole genome amplification. Genomic DNA was spiked at six concentrations for Illumina sequencing (Table 1). We identified *B. anthracis* DNA by mapping sequenced reads to a defined subset of reference genomes. The target reference set included the *B. anthracis* Ames complete genome as well as the pXO1 and pXO2 virulence plasmids. The background reference set included eleven finished genomes chosen to represent environmental background and evaluate specificity and sample variation (Table 2).

The relative number of reads mapping to the target and background reference genomes using *B. anthracis*-spiked aerosol and soil samples were normalized to the total number of reads obtained for each sample (Figure 1A-B). Absolute read numbers are shown in Supplementary Figure S1A-B. The relative number of Illumina reads in each aerosol sample mapping to the *B. anthracis* chromosome and plasmids increased proportionately from the samples containing one genomic copy to the samples containing 100,000 copies (Figure 1A). There was no appreciable increase in the number of corresponding hits to the select set of background species. The number of reads mapping to *B. anthracis* was greater than those mapping to background genomes when as few as 10 genomic copies were present in aerosol samples (100 pg aerosol DNA). We observed similar patterns in soil samples, as the percentage of reads mapped to target genomes increased with increasing genome copy number

(Figure 1B). The percentage of reads mapping to *B. anthracis* exceeded those mapping to background genomes when as few as 100 genomic copies were present in soil samples (1 ng soil DNA), indicating a limit of detection of 100 copies/ng background DNA.

Parallel detection of B. anthracis *in environmental samples by 454 sequencing*

Comparable trends were observed when processing samples using 454 deep sequencing. *B. anthracis* genomic DNA was spiked at four concentrations (Table 1) and sequencing data aligned to reference genomes (Table 2). The percentage of reads mapping to *B. anthracis* from both aerosol and soil samples increased with increasing genomic copy number without an observable trend in the number of matches to select background species (Figure 1C-D). Absolute read quantities are given in Supplementary Figure S1C-D. Similar to our Illumina data, the percentages of reads mapping to *B. anthracis* target genomes in both aerosol and soil samples were markedly higher than those mapping to background for samples containing as few as 10 genomic copies (Figure 1C-D). These data indicate that 454 sequencing is capable of detecting as few at 100 bacilli/ng aerosol background DNA and 10 bacilli/ng soil background DNA.

Sequencing via 454 yielded higher signal relative to noise when comparing the ratio of reads mapping to target versus reads mapping to background genomes (Figure 1E). We also observed superior signal ratios when sequencing material from aerosol backgrounds compared to soil backgrounds, although this is likely due to the reduced quantity of background DNA (100 pg compared to 1 ng) used in aerosol relative to soil samples. Overall, our combined data from Illumina and 454 sequencing indicate accurate detection of *B. anthracis* by deep sequencing when as few as 10-100 genomic copies are present in a complex environmental sample.

Specificity of B. anthracis *Ames detection by Illumina sequencing*

Following our analysis of limit of detection, we sought to determine the specificity of sequencing detection. We calculated the percentage of reads from sequencing runs mapping to the *B. anthracis* Ames genome as well as the pXO1 and pXO2 plasmids. These data were compared to the percentage of reads mapping to two close relatives. The strains examined for this purpose were *B. thuringiensis* Al Hakam, including the pALH1 plasmid, and *B. cereus* biovar anthracis strain CI, including the pCI-X01 and pCI-X02 plasmids.

In both the soil and aerosol background samples, the proportion of reads mapping to *B. anthracis* Ames was consistently higher than the proportion of reads mapped to *B. thuringiensis* or *B. cereus* (Figure 2). The difference in proportion of mapped reads between *B. anthracis* and close species, however, was not highly significant. For aerosol samples, across multiple genome copy numbers, we observed a 1.30 ± 0.02 fold increase in the proportion of reads mapping to *B. anthracis* relative to *B. thuringiensis*, and a 1.18 ± 0.01 fold increase relative to *B. cereus*. For soil samples, we similarly observed 1.22 ± 0.11 and 1.14 ± 0.07 fold increases respectively. Such narrow distinctions are not adequate for specific detection. The ability to discriminate between strains was, therefore, limited when comparing all mapped reads, due to a high degree of sequence similarity between *B. anthracis* and its close relatives. We therefore employed a modified approach in which only reads mapping to one of the organisms (*B. anthracis* or the close relative) were included; reads that mapped to both or neither reference genomes were not considered. The log value of the number of uniquely mapped Illumina reads obtained for each *Bacillus* species with increasing copy number is plotted in Figure 3. We observed a greater than 1000-fold increase in the number of reads uniquely mapping to *B. anthracis* compared to *B. thuringiensis* (Figure 3A-B) and a greater than 300-fold increase when compared to *B. cereus* for each sample (Figure 3C-D). It is therefore possible, using this unique mapping approach, to differentiate *B. anthracis* from closely related *Bacillus* species. These data indicate that Illumina sequencing is capable of identifying *B. anthracis* in environmental samples despite cross-mapping of large read quantities to very similar *Bacillus* species.

Specificity of B. anthracis *Ames detection by 454 sequencing*

Specificity of 454 sequencing was also examined by mapping reads to *B. anthracis* Ames and pXO plasmids as well as the closely related *B. thuringiensis* and *B. cereus* and their respective plasmids. Again, given our experience with interpretation of the Illumina short read data, sequencing reads mapping to multiple strains were not considered; only uniquely mapped reads were counted. We observed parallel results to those obtained by Illumina sequencing, with a higher proportion of reads mapping to *B. anthracis* Ames than its near neighbors (Figure 4). When 100 *B. anthracis* copies were included in environmental samples, the number of reads uniquely mapping to *B. anthracis* exceeded those mapping to *B. thuringiensis* by greater than 300-fold, and those mapping to *B. cereus* by greater than 100-fold (in both aerosol and soil backgrounds). These results confirm our previous observation that deep sequencing is a robust method for the specific detection of *B. anthracis*, and is capable of making accurate distinctions from other closely related species.

GenBank BLAST analysis of Illumina short read sequences

In our final sequencing analysis, we tested a more comprehensive detection strategy by mapping sequencing data to the full NCBI GenBank database, creating a taxonomic distribution of all organisms in each sample. For Illumina reads, a local version of the database was used, and only GenBank sequences classified as bacteria or Archaea were considered. We compiled the number of uniquely mapped reads and identified the 15 taxa with the most hits in each sequencing run, creating a union set of top hits across all samples (22 unique species in aerosol and 24 unique species in soil samples) (Figure 5A-B). No additional normalization was necessary due to filtering steps accounting for major sources of expected bias. In the aerosol background, *B. anthracis* was identified as a top 10 hit when 10 or more genomic copies were present, and was the number one hit with 1000 or more copies present. In the soil background, 100 or more genomic copies were required in order for *B. anthracis* to be detected in the top 10, with 10,000 or more copies needed in order to reach the number one position, indicating a 10-100 genomic copy limit of detection when screening across the entire Genbank reference. Other organisms commonly observed in aerosol backgrounds included *Ralstonia pickettii*, *Cupriavidus metallidurans*, and *Delftia acidovorans*. Soil samples also exhibited *R. pickettii* and *C. metallidurans*, in addition to *Nitrosospira multiformis* (Supplementary tables S1-S2).

Megablast analysis of 454 sequencing data

We also performed a parallel whole microbiome analysis of 454 sequencing data. We compiled the 25 taxa with the highest number of uniquely mapped reads in each sample and created a union set of top hits (Figure 5C-D). Due to the broad variation in number of sequencing reads generated for each sample, we normalized the occurrences of each organism to the total number of reads in the sample. These results show similar patterns to those observed in the Illumina approach above, with an approximately 10-fold increase in the number of megablast hits to *B. anthracis* and the related *B. cereus* genomes with increasing copy number. As expected, other organisms commonly observed in these samples included *Ralstonia* and *Cupriavidus*. *B. anthracis* remained a highly observed sequencing hit, even when processing the environmental background across the entire GenBank reference database.

Determination of B. anthracis *Ames limit of detection in aerosol and soil samples by microarray*

Finally, we performed detection testing using a census microbial microarray developed at LLNL, including probes designed for both census and detection purposes. We designed detection probes to be conserved across multiple sequences from within a family, but not across families or kingdoms. Such probes aim to detect known organisms or discover novel organisms exhibiting some homology to species which have been previously sequenced, particularly in those regions known to be conserved. We previously designed the Lawrence Livermore Microbial Detection Array (LLMDA) using this approach [12]. Census probes, in contrast, represent the least conserved regions, and are the most strain or isolate specific probes. Such census probes aim to provide higher level discrimination and identification of known species and strains to facilitate forensic resolution.

We tested the census array using the same serially diluted *B. anthracis* Ames genomic DNA spiked into either aerosol or soil samples. We fluorescently labeled DNA and hybridized each sample to the census array in duplicate. Analysis was performed using algorithms designed at LLNL, which identified species most likely to be present in each sample based on hybridization intensity. We observed that, in the presence of 100 pg aerosol background DNA, 100 copies of the *B. anthracis* Ames genome were required for successful microarray detection in both replicates (Table 3). In the presence of 1 ng soil background DNA, 1000 genome copies were required for detection. These results indicate a limit of detection of 1000 bacilli/ng background DNA for the census microarray. A summary of the detection limits for each method employed in this study is given in Table 4.

DISCUSSION

The life cycle and virulence of *B. anthracis* distinguish it as an organism with a high potential for use as an agent of bioterrorism. Vegetative spores are highly resistant to extreme, conventionally bactericidal conditions and can be easily dispersed. The acute and potentially fatal nature of anthrax disease poses a significant threat to human health in the context of a broad exposure incident. The attacks of 2001 raised awareness of the hazards resulting from *B. anthracis* release and underscored the importance of biosurveillance. A vital component of the United States' effort toward the prevention of such an attack is the ability to detect *B. anthracis* in an environmental sample. These efforts present multiple difficulties, including the presence of extraneous contaminants, low bacterial concentrations, and non-cultivable bacilli [13]. We demonstrated in this study that next-generation sequencing represents a sensitive and specific technique for identification of *B. anthracis*, capable of overcoming many of these obstacles.

The increasing availability and falling cost of next-generation sequencing provides the opportunity to perform whole genome analyses of pathogenic microbes, providing a breadth of information not available from more limited and focused genetic protocols such as PCR or Sanger sequencing. High-throughput sequencing has been used to characterize isolates of *B. anthracis* and successfully identified SNPs corresponding to distinct strains [14]. The Amerithrax investigation of the 2001 anthrax attacks used whole genome sequencing and comparative analyses to identify unique genomic characteristics of the *B. anthracis* strains sent in public letters, validating the forensic potential of this technology [8, 15].

Detection by next-generation sequencing has previously been evaluated using purified versions of isolated *Bacillus* strains [16, 17], and use of deep sequencing has proven capable of identifying unique and minute genetic characteristics in *B. anthracis* [17]. Processing of biothreat organisms via sequencing had not previously, however, been validated in the context of an environmental background [8]. Since these are the samples likely to be encountered in the event of bacillary exposure, we performed screening for *B. anthracis* genomic DNA in both aerosol and soil backgrounds, using multiple genome copy numbers to identify limit of detection thresholds. We evaluated samples on two separate sequencing platforms, Illumina and 454. Detection at each threshold was subsequently analyzed for specificity by comparison to *B. cereus* and *B. thuringiensis*.

We analyzed the results from each sample using two different procedures. The first method used Burrows-Wheeler Aligner (Illumina reads) and gsMapper (454 reads) software to map reads to specific known genomes, and the second used Bowtie (Illumina reads) and megablast (454 reads) to query the NCBI GenBank database. The mapping approach using a defined number of reference genomes was much faster, with runtimes of several minutes compared to hours for the more comprehensive NCBI queries. A megablast analysis, however, will, of course, provide a more comprehensive approximation of all known organisms in a given sample. Both methods gave similar results in the number of matches to *B. anthracis*, with a limit of detection of 100 copies/ng environmental DNA.

The full GenBank analysis revealed a number of additional species present in environmental samples, particularly those belonging to the genera *Ralstonia* and *Cupriavidus*. Each of these species are adapted to survival in terrestrial, soil environments, and are not known to be pathogenic. They represent microorganisms likely to be observed as background when comparable analyses of environmental samples are performed. It is therefore important to note that their presence does not adversely affect our ability to detect *B. anthracis*.

One of the major challenges to specific detection of *B. anthracis* is its similarity to both *B. cereus* and *B. thuringiensis*. Some sequence analyses have indicated *B. anthracis* should be classified as a lineage of *B. cereus*, with the primary distinction being differing plasmids [18]. Even these distinctions, however, may be dependent on strain origin, as theoretically non-pathogenic *Bacillus* species have been observed to cause disease and harbor virulence genes typical of *B. anthracis* [19-21]. It has been suggested, therefore, that the uniqueness of *B. anthracis* is actually a more complex product of co-evolution of the genome with its corresponding plasmids [11, 20]. Similarly, within *B. anthracis*, individual strains share a tremendous degree of similarity, leading to the suggestion that *B. anthracis* represent the most genetically homogenous known bacteria [11]. All of the above contribute challenges to specific identification of *B. anthracis* via nucleic acid-based methodology.

We hypothesized that the comprehensive nature of sequencing would yield heightened specificity in making these difficult distinctions. When mapping to defined genomes for Illumina and 454 data, in both aerosol and soil samples, we observed a greater than 300 fold amplification in number of reads mapping uniquely to *B. anthracis* compared to its near neighbors. We did observe lower specificity when mapping to the full NCBI database. It is possible that, when using the NCBI database, the availability of a more comprehensive near neighbor pan-genome contributed to the observed differences. Additionally, it must be taken into account that only uniquely mapped reads were counted in the Burrows Wheeler/gsMapper analysis. Due to lower specificity and greater background variability, the detection reliability at very low concentrations was not as high in the NCBI GenBank approach. Taken in full, these data suggest that our sequencing approach and unique combination of mapping strategies could be a highly effective way to approach the problematic task of distinguishing *B. anthracis* strains from other members of the *B. cereus* group.

These data analyses also provide broader insight into interference in mapping classification caused by sequence similarity between bacterial genomes, particularly when mapping short reads in closely-related but distinct bacterial species. The use of reads that uniquely map to a given species is an important step toward remedying the detection of false-positives during microbiome characterization. One potential future difficulty in our approach lies in the fact that, without prior information, it may be difficult to estimate the exact quantity of *B. anthracis* in environmental samples via sequencing, as this would be a function of the proportion of total reads mapping to the target database. The assumption that the background is identical in size, quantity, and diversity in all samples may not be possible to confirm. One alternative could be to extensively over-sequence the sample, so that read saturation is not a significant confounding factor in estimation of the amount of target organism in the sample.

In order to assess an alternative detection technology, we used a comprehensive microbial census array developed at LLNL. We included two classes of probes on the array, census and detection, to maximize the capacity for identification of well-characterized as well as novel microbes and to facilitate high confidence at multiple taxonomic levels. We determined the array limit of detection for *B. anthracis* to be 1000 copies/ng background DNA. Our observed sensitivity was lower than that of the next-generation sequencing approach; however, the array provides a platform which is more cost-effective and higher-throughput than sequencing, while still maintaining comprehensive target specificity. This microarray represents a complementary screening technology for biothreat detection, where subsequent sequencing can be performed if it is determined that heightened genomic detail and specificity are required. It should be noted that the application described in this study demonstrates only a small proportion of the capabilities of the census array, and that this technology exhibits far-reaching potential in many fields of microbial analysis, including surveillance, forensics, and clinical characterization.

Studies examining sensitivity limits for *B. anthracis* have been performed in the past. Many of these experiments have been effectively summarized in several reviews [8, 11, 13]. Methods applied include RT-PCR, microarray, ELISA-based immunoassays, spectroscopy, mass spectrometry, biosensor assays, and high-throughput sequencing. Comparison of detection limits is highly dependent on sample medium. We chose to employ a background medium of environmental DNA, as this would be the primary confounding factor for specific detection from high-throughput sequencing. Reported limits of detection in soil have ranged from 0.1 CFU/g soil (PCR-based) to 3.2 x 10^8 CFU/g soil (fiber optic biosensor assay) [22, 23]. Although relatively few studies have been performed employing an aerosol background, limits of detection have been reported as low as 35-39 spores/m^3 air using an aerodynamic particle sizer [24]. A more specific ELISA biochip assay used to screen air samples for *Bacillus* reported higher thresholds, at 17 CFU/liter air [25]. Comparison of our data to these results is complicated by our quantification of background DNA in our detection calculations.

Finally, it is important to note that the open potential of sequencing to bring to light new or engineered variants of *B. anthracis* is an important feature of this detection method. Increasing availability of sequencing technology and its amplified presence in laboratories worldwide should make next-generation sequencing an important component of the biodefense surveillance strategy moving forward, particularly in cases where anamolous and emergent strains are a concern.

MATERIALS AND METHODS

DNA Extraction from environmental samples

We collected soil in the downtown areas of both Oakland and San Francisco, California, collecting four samples in each city at multiple sites. We extracted nucleic acid using the UltraClean Soil DNA Isolation Kit (MoBio) using the manufacturer's alternative protocol for maximum yield. Following extraction, we used 1 ng of each extracted DNA in a real-time PCR assay to test for inhibition. All samples showed a high level of PCR inhibition, thus we reprocessed DNA starting from Step 12 of the MoBio alternative protocol, intended to remove excess humic acid. We obtained primary sampling filters from BioWatch aerosol collection units (collected April, 2009) from the National Capital Region Laboratory. We processed filters and extracted DNA as described previously (Paper in preparation) and quantified DNA concentrations using the Qubit fluorometer (Invitrogen).

Addition of Bacillus anthracis *Ames DNA to environmental samples*

We acquired *B. anthracis* Ames DNA from the select agent laboratory within LLNL, confirming sterility by plating on blood agar. We quantified DNA and determined copy number using the Qubit fluorometer (Invitrogen). We added genomic DNA to environmental nucleic acid at six concentrations for Illumina sequencing (1, 10, 100, 1000, 10000, and 100000 genomic copies) and four different concentrations for 454 sequencing (0, 1, 10, and 100 genomic copies). We mixed each sample with 100 pg of extracted DNA from aerosol filters or 1 ng DNA from a combination of Oakland and San Francisco soil extracts.

Whole genome amplification and purification

We amplified *B. anthracis*-spiked samples using the REPLI-g Midi Kit (Qiagen), intended to provide uniform whole genome amplification. This kit was used to amplify each copy number dilution of *B. anthracis* DNA spiked in either 1 ng soil DNA or 100 pg aerosol DNA, according to the manufacturer's instructions, allowing samples to amplify for 16 hours at 30°C. We purified amplified samples using the Qiaquick PCR Purification Kit (Qiagen).

Illumina and 454 sequence generation

We provided amplified samples containing 1, 10, 100, 1000, 10000, or 100000 copy numbers of the *B. anthracis* genome to Eureka Genomics for sequencing using the Illumina platform. Eureka Genomics prepared a paired end non-indexed standard Illumina library for each sample and generated sequencing reads using Illumina GAIIx sequencing technology, running one sample per lane. Soil samples were sequenced with 51 cycles of paired end reads, while aerosol samples were sequenced with 51 cycles of single end reads. For the purposes of analysis, the paired end reads were decoupled and used as if single reads were generated. As a result, soil samples exhibit roughly double the number of sequence reads (decoupled singletons). It should be noted that while the number of sequencing reads in the soil data is doubled, the number of independent samplings is not. Since the paired end read is typically generated from a single sequence fragment, the sensitivity of the soil samples, in terms of the ability to detect rare environmental organisms, is not significantly increased when compared to aerosol samples. We provided soil and aerosol samples containing 0, 1, 10, or 100 genome copies to Brigham Young University (BYU), where the bioinformatics department generated 454 sequencing reads from one 96-well plate using recommended protocols (454 Life Sciences, Roche).

Illumina sequence mapping and data analysis

Eureka Genomics performed analysis on reads generated by Illumina sequencing. We used publicly available software (Burrows-Wheeler Aligner) for mapping of Illumina reads to specified reference genomes. For the Illumina-based whole microbiome mapping approach, the length of Illumina GAIIx reads (36 and 51 bp) resulted in a high likelihood of multiple optimal alignments in different genomes. As such, a Megablast top-hit only approach would carry an amplified risk of producing an excessive false positive hit rate, and would be prohibitively computationally expensive. We therefore employed publically available short-read mapping software (Bowtie) instead of Megablast, keeping all hits up to three mismatches. We parsed the resulting output from each Bowtie run to obtain taxonomy IDs (taxID) matched by each read. All possible hits for each read were recorded and classified using NCBI.

It is important to note that multiple sub-strain reference genomes can artificially inflate the number of mapped reads for a given species. In order to avoid such bias, as well as inflation associated with tandem repeats, each read was counted as matching to a given taxID only once. Additionally, we collapsed bacterial strains such that reads mapping to sub-strains were instead counted as mapping to parent species. This is particularly helpful when trying to distinguish between reads mapping to closely related species that share significant sequence similarity, as mapping to multiple sub-strains of a given bacterial species could erroneously suggest that the higher level species is present at low levels in the sample.

454 sequence mapping and data analysis

Oregon Ridge National Laboratory (ORNL) performed analysis on the 454 sequencing reads generated by BYU. We used the vendor-provided gsMapper software (Roche Scientific) to map 454 reads to corresponding reference sequences using a 98% minimum identity cutoff. For the whole microbiome analysis of 454 reads, we used the megablast program from NCBI version 2.2.18, with nucleotide and taxonomy databases as of December, 2010. We split each input file into 256 smaller files, using each of these smaller files as input for megablast, running on a local Linux cluster. We parsed the resulting output from each megablast run to obtain the best hit NCBI sequence identification and corresponding bacterial species for each queried sequence.

Census microarray probe design

We downloaded all available bacterial and viral genome and fragment sequences from NCBI GenBank, the Joint Genome Institute, the J. Craig Venter Institute, the Sanger Institute, and additional proprietary whole-genome databases from collaborators, including the San Francisco Blood Systems Research Institute. Sequence data for complete genomes, viral segments, and plasmids were current as of August 2009, and for sequence fragments as of January 2009. We began the probe design process by identifying family specific sequence regions (Supplementary Table S3). For each target family we eliminated regions with perfect matches to sequences outside the target family. Using the suffix array software vmatch [26], perfect match subsequences of at least 17 nt long present in non-target viral families or 25 nt long present in the human genome or non-target bacterial families were eliminated from consideration as possible probe subsequences.

From these family-specific regions, we designed probes 50-66 bases long using previously described methods [27]. Briefly, we generated candidate probes using Primer3 [28], followed by T_m and homodimer, hairpin, and probe-target free energy (ΔG) prediction using Unafold [29]. Candidate probes with unsuitable ΔG or T_m were excluded as described previously [27]. The range for these parameters included length of 50-66 bp, $T_m \geq 80^{\circ}C$, GC% 25-75%, ΔG of homodimer formation > 15 kcal/mol, ΔG of hairpin formation > -11 kcal/mol, and $\Delta G_{\text{adjusted}}$ $(\Delta G_{complement} - 1.45\ \Delta G_{hairpin} - 0.33\ \Delta G_{homodimer}) \leq -52$ kcal/mol. An additional minimum sequence complexity constraint was enforced, requiring a trimer frequency entropy of at least 4.5. In the event that an insufficient number of candidate probes per target sequence passed all criteria, parameters (first Unafold and then Primer3) were relaxed to allow an adequate number of probes per target. We then BLASTed candidate probes against the family of target sequences from which they were designed to identify which sequences should be represented by each candidate. A target was considered to be represented if it matched a probe with > 85% sequence similarity over the total probe length, with a 29 contiguous base perfect match spanning the central probe base. We then ranked probe candidates by conservation level.

Separate strategies were employed for detection and census probes. Detection probes were selected as previously described [12]. Essentially, probes corresponding to a higher number of targets in the family were chosen preferentially. Conversely, for census probes, probes detecting fewer targets in the family were chosen preferentially. A secondary dispersal ranking was used to favor probes with genomic loci distant from those which had already been selected to represent the target. We included 5-30 detection probes per target sequence and 1-10 census probes per target depending on the array density. Approximately 1,000-3,500 random negative control sequences, with matched length and GC content, were included. These controls had no appreciable homology to known sequences based on BLAST similarity, and were used to assess background hybridization intensity. The standard array used in this study was designed to fit the NimbleGen 388K format (Roche).

Microarray hybridization and data analysis

Amplified environmental DNA spiked with *B. anthracis* Ames genomic DNA was fluorescently labeled using the NimbleGen One-Color DNA Labeling Kit (Roche) according to recommended protocols. DNA was purified and hybridized to the census array using the NimbleGen hybridization kit. Samples were allowed to hybridize for 17 hours and washed using the NimbleGen Wash Buffer Kit (Roche). Microarrays were scanned on the Axon GenePix 4000B 5 µm scanner (Molecular Devices). Image files were aligned using NimbleScan (Version 2.4) software, and pair text files were exported for data analysis. A previously described maximum likelihood analysis method was used to analyze the microbial hits from samples hybridized to the array [12].

ACKNOWLEDGEMENTS

This work was funded by the Department of Homeland Security and performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

CONFLICT OF INTEREST STATEMENT

The authors do not have a commercial or other association that might pose a conflict of interest.

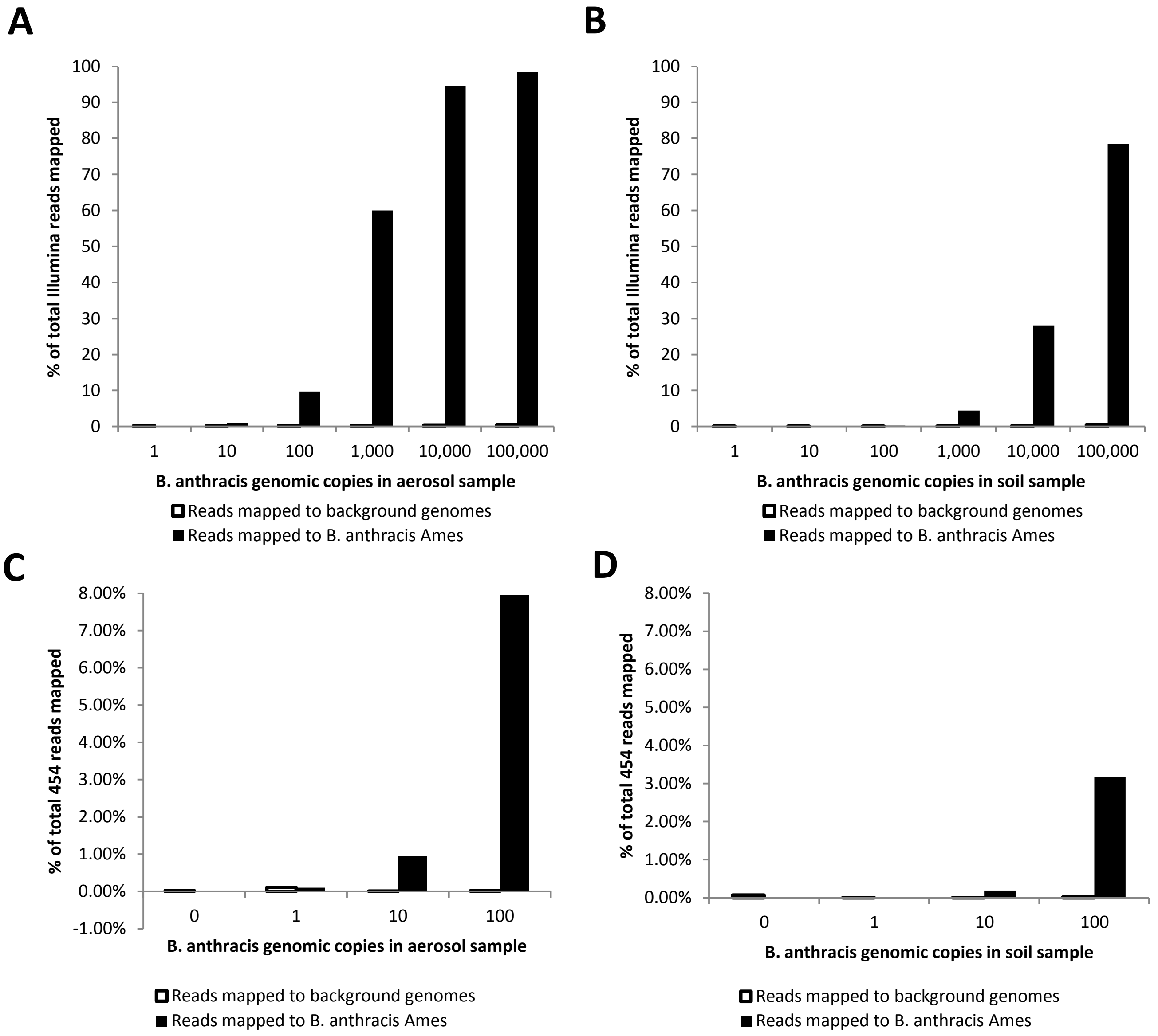
Table 1. *B. anthracis* genome copy numbers added to environmental background DNA for detection assessment.

***B. anthracis* copy #**	100,000	10,000	1,000	100	10	1
			Aerosol background			
Amount *B. anthracis* DNA	560 pg	56 pg	5.6 pg	560 fg	56 fg	5.6 fg
Amount aerosol filter DNA	100 pg	100 pg	100 pg	100 pg	100 pg	100 pg
% *B. anthracis* DNA in aerosol DNA	98.20%	35.90%	5.30%	0.56%	0.060%	0.006%
			Soil background			
Amount *B. anthracis* DNA	560 pg	56 pg	5.6 pg	560 fg	56 fg	5.6 fg
Amount soil DNA	1 ng	1 ng	1 ng	1 ng	1 ng	1 ng
% *B. anthracis* DNA in soil DNA	35.90%	5.30%	0.56%	0.060%	0.006%	0.001%

Table 2. Bacterial reference genomes used for mapping of Illumina and 454 sequencing reads.

Target Reference Genomes
Bacillus anthracis str. Ames, complete genome
Bacillus anthracis virulence plasmid PX01, complete sequence
Bacillus anthracis plasmid pXO2, complete sequence
Bacillus thuringiensis str. Al Hakam, complete genome
Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence
Used as reference for 454 sequencing reads only
Bacillus anthracis str. Sterne, complete genome
Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence
Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence
Used as reference for Illumina sequencing reads only
Bacillus cereus biovar anthracis str. Cl, complete genome
Bacillus cereus biovar anthracis str. Cl plasmid pCl-XO1, complete sequence
Bacillus cereus biovar anthracis str. Cl plasmid pCl-XO2, complete sequence
Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14, complete sequence
Background Reference Genomes
Burkholderia pseudomallei strain K96243, chromosome 1, complete sequence
Escherichia coli O157:H7 EDL933, complete genome
Francisella tularensis subsp. tularensis SCHU S4 complete genome
Pseudomonas aeruginosa PAO1, complete genome
Rhodopseudomonas palustris CGA009 complete genome
Sinorhizobium meliloti 1021 complete chromosome
Staphylococcus aureus subsp. aureus N315 DNA, complete genome
Streptomyces coelicolor A3 (2) complete genome
Yersinia pestis CO92 complete genome
Bacillus subtilis subsp. subtilis str. 168 complete genome
Clostridium botulinum A str. Hall, complete genome

Figure 1. Mapping of sequencing reads obtained from *B. anthracis*-spiked environmental samples to specified reference genomes. We combined *B. anthracis* Ames genomic DNA with background nucleic acid extracted from either aerosol or soil-based material. We spiked increasing genome copy numbers into samples at 10-fold concentration intervals. We then subjected these samples to both Illumina and 454 deep sequencing, and mapped resultant reads to either a target set (*B. anthracis*) or a background set, comprising bacterial species likely to be found in an environmental background. In order to standardize our results, numbers of reads mapped were normalized to total reads obtained for each sample. Shown are percent of reads mapped for **A.** Illumina reads from aerosol background, **B.** Illumina reads from soil background, **C.** 454 reads from aerosol background, and **D.** 454 reads from soil background. **E.** Signal to noise ratios are given by dividing number of reads mapped to the target reference by the number of reads mapping to the background reference.


A
% of total Illumina reads mapped
100
90
80
70
60
50
40
30
20
10
0
1
10
100
1,000
10,000
100,000
B. anthracis genomic copies in aerosol sample
□ Reads mapped to background genomes
■ Reads mapped to B. anthracis Ames
B
% of total Illumina reads mapped
100
90
80
70
60
50
40
30
20
10
0
1
10
100
1,000
10,000
100,000
B. anthracis genomic copies in soil sample
□ Reads mapped to background genomes
■ Reads mapped to B. anthracis Ames
C
% of total 454 reads mapped
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
-1.00%
0
1
10
100
B. anthracis genomic copies in aerosol sample
□ Reads mapped to background genomes
■ Reads mapped to B. anthracis Ames
D
% of total 454 reads mapped
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
0
1
10
100
B. anthracis genomic copies in soil sample
□ Reads mapped to background genomes
■ Reads mapped to B. anthracis Ames

E

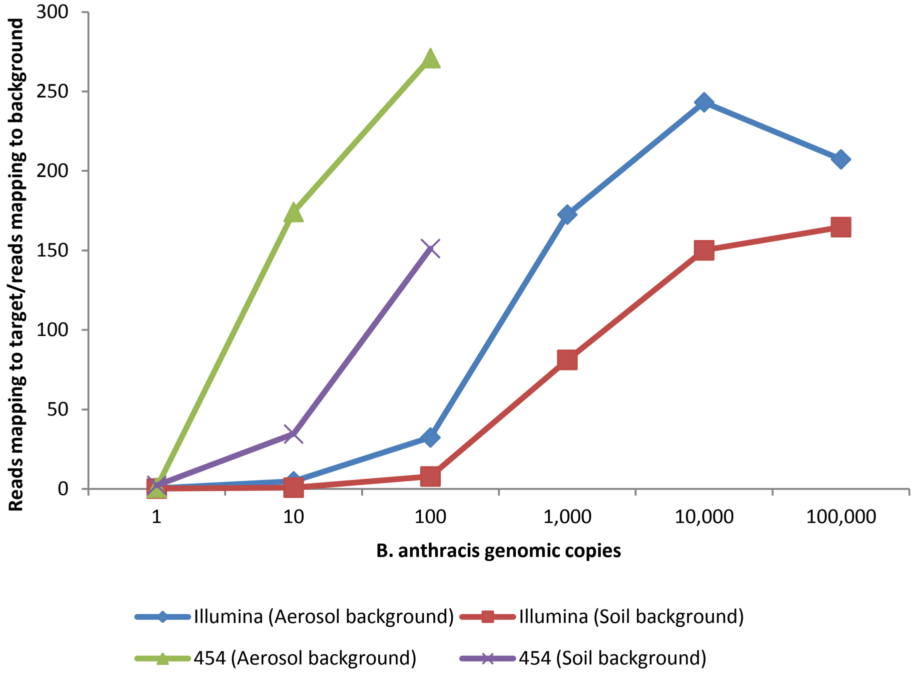

300
250
200
150
100
50
0
Reads mapping to target/reads mapping to background
1
10
100
1,000
10,000
100,000
B. anthracis genomic copies
Illumina (Aerosol background)
Illumina (Soil background)
454 (Aerosol background)
454 (Soil background)

Figure 2. Correspondence of Illumina reads to closely related *Bacillus* species. Following sequencing of *B. anthracis*-spiked environmental samples, we proceeded to identify the specificity of this approach to *B. anthracis* detection. We examined mapping specificity by determining the percent of total Illumina reads mapping to the closely related species *B. thuringiensis* Al Hakam and *B. cereus* biovar anthracis CI in genetic material extracted from **A.** aerosol and **B.** soil background samples.

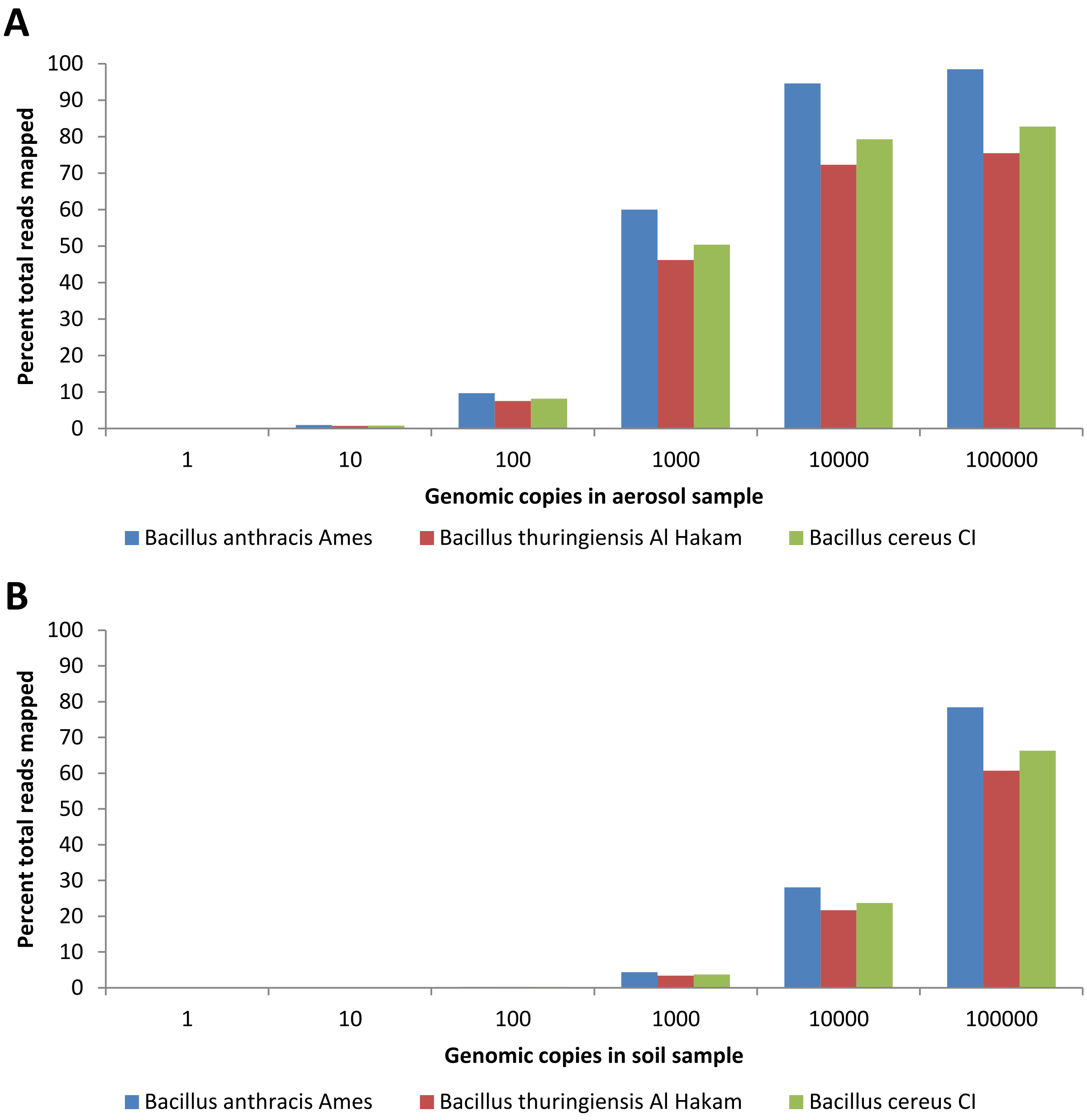

A
Percent total reads mapped
0
10
20
30
40
50
60
70
80
90
100
1
10
100
1000
10000
100000
Genomic copies in aerosol sample
Bacillus anthracis Ames
Bacillus thuringiensis Al Hakam
Bacillus cereus CI
B
Percent total reads mapped
0
10
20
30
40
50
60
70
80
90
100
1
10
100
1000
10000
100000
Genomic copies in soil sample
Bacillus anthracis Ames
Bacillus thuringiensis Al Hakam
Bacillus cereus CI

Figure 3. Alignment of uniquely mapped Illumina reads to *B. anthracis* and closely related species. Due to the high degree of sequence similarity between the examined *Bacillus* species, we elected to use a unique mapping approach. We identified reads mapping only to *B. anthracis* or a near neighbor; reads mapping to multiple reference genomes were discarded. This approach facilitated distinction between species. **A.** Log sequencing reads mapping uniquely to *B. anthracis* or *B. thuringiensis* in aerosol samples. **B.** Log reads mapping uniquely to *B. anthracis* or *B. thuringiensis* in soil samples. **C.** Log sequencing reads mapping uniquely to *B. anthracis* or *B. cereus* in aerosol samples. **B.** Log reads mapping uniquely to *B. anthracis* or *B. cereus* in soil samples.

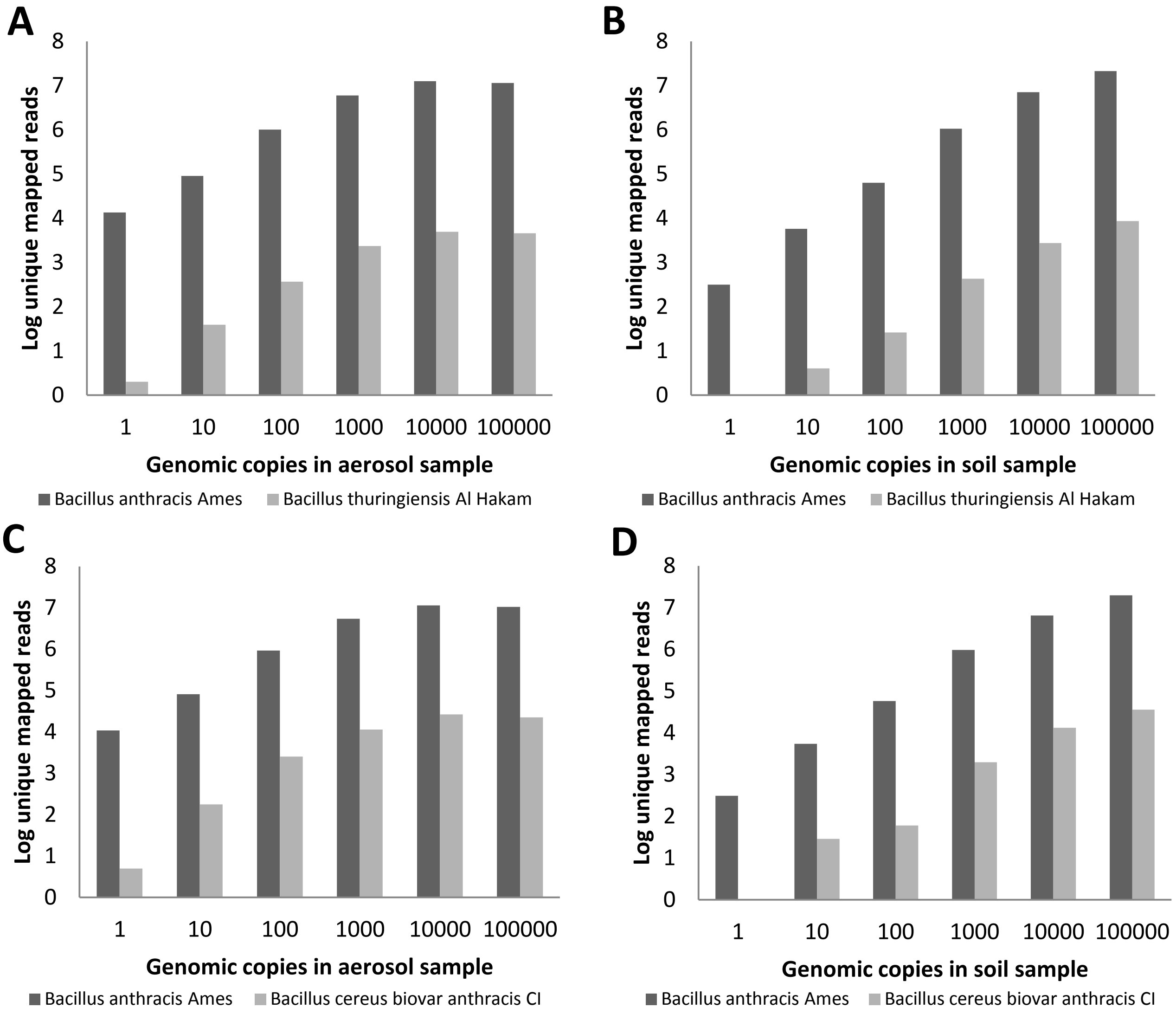

A
B
C
D
Log unique mapped reads
0
1
2
3
4
5
6
7
8
1
10
100
1000
10000
100000
Genomic copies in aerosol sample
Genomic copies in soil sample
Bacillus anthracis Ames
Bacillus thuringiensis Al Hakam
Bacillus cereus biovar anthracis CI

Figure 4. Alignment of unique 454 reads to *B. anthracis* and near-neighbor species. As in figure 3, we identified 454 reads mapping only to *B. anthracis* or a close relative, discounting reads mapping to multiple reference genomes. Similar to our observations with Illumina reads, this approach facilitated distinction between species. **A.** Log sequencing reads mapping uniquely to *B. anthracis* or *B. thuringiensis* in aerosol samples. **B.** Log reads mapping uniquely to *B. anthracis* or *B. thuringiensis* in soil samples. **C.** Log sequencing reads mapping uniquely to *B. anthracis* or *B. cereus* in aerosol samples. **B.** Log reads mapping uniquely to *B. anthracis* or *B. cereus* in soil samples.

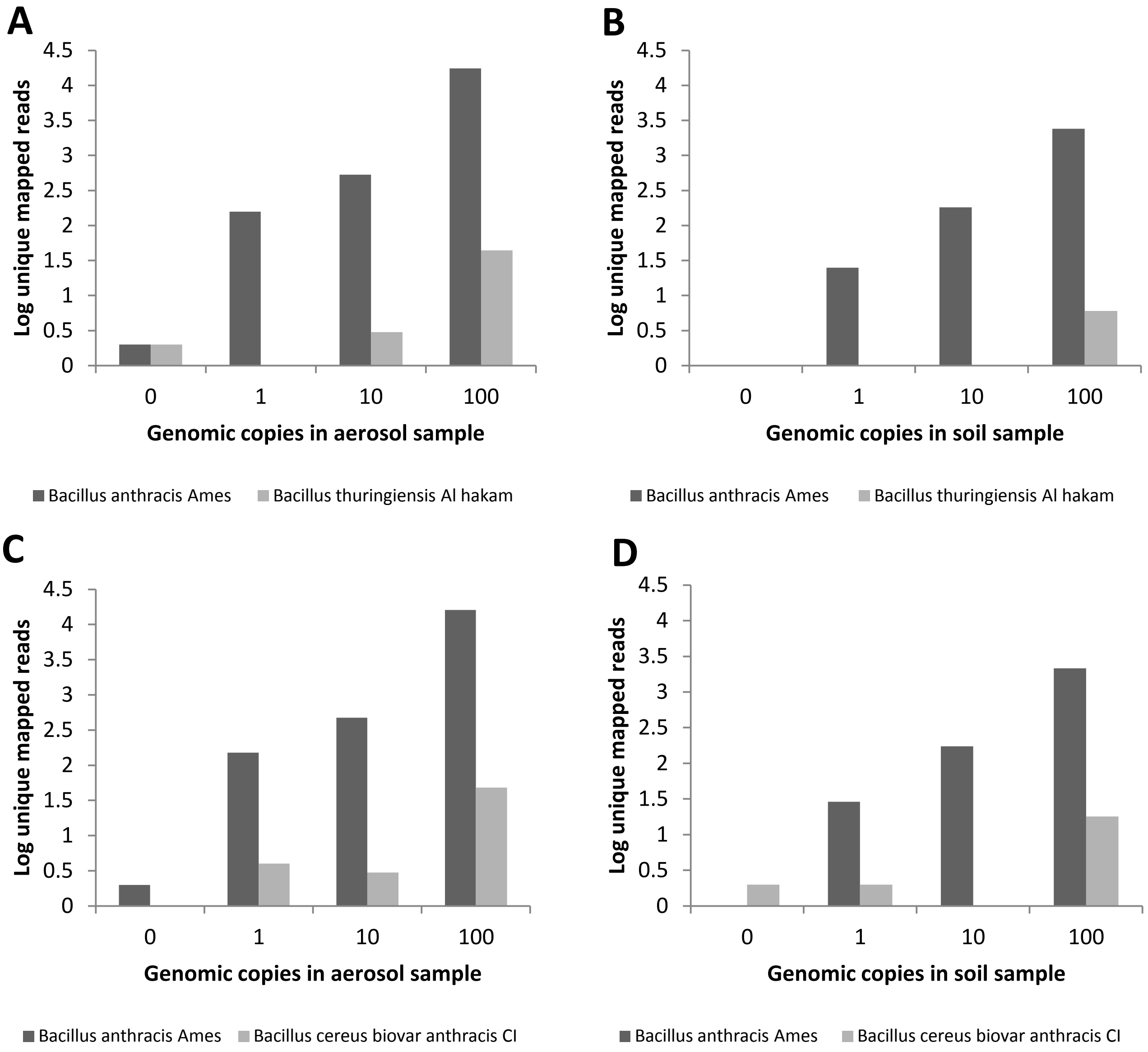
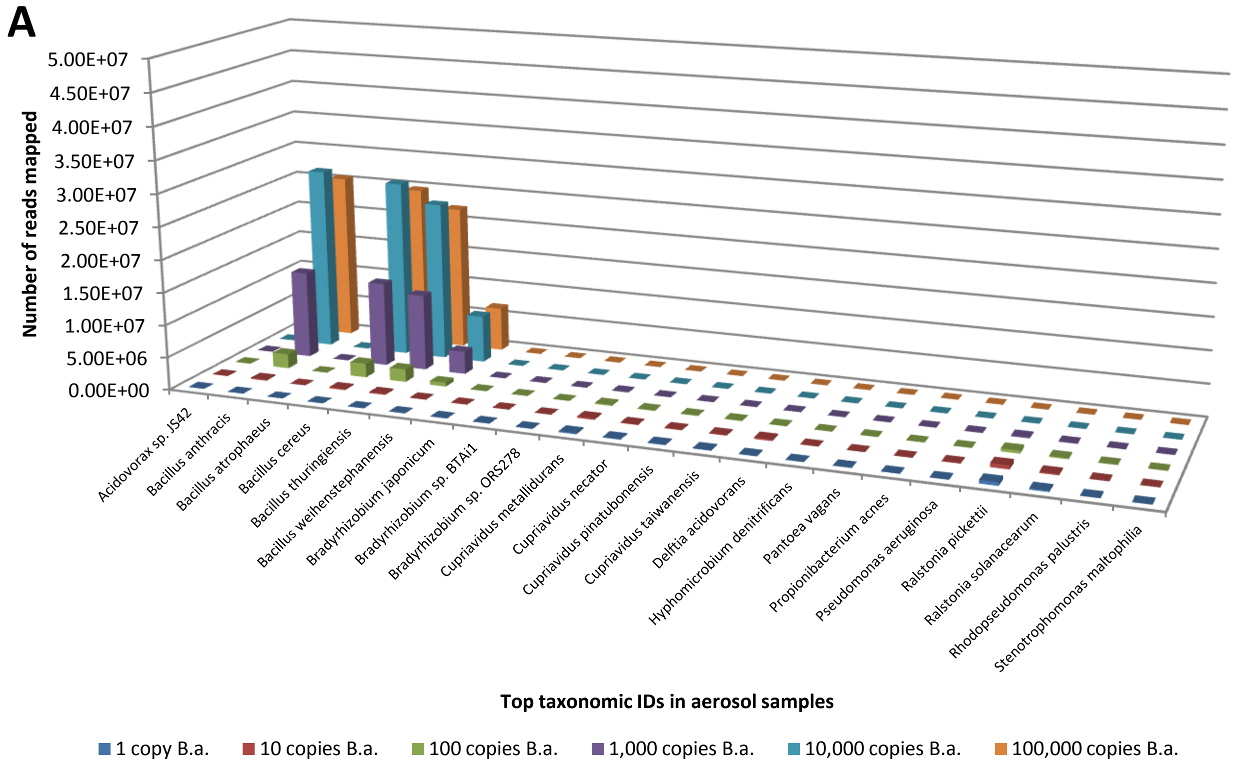
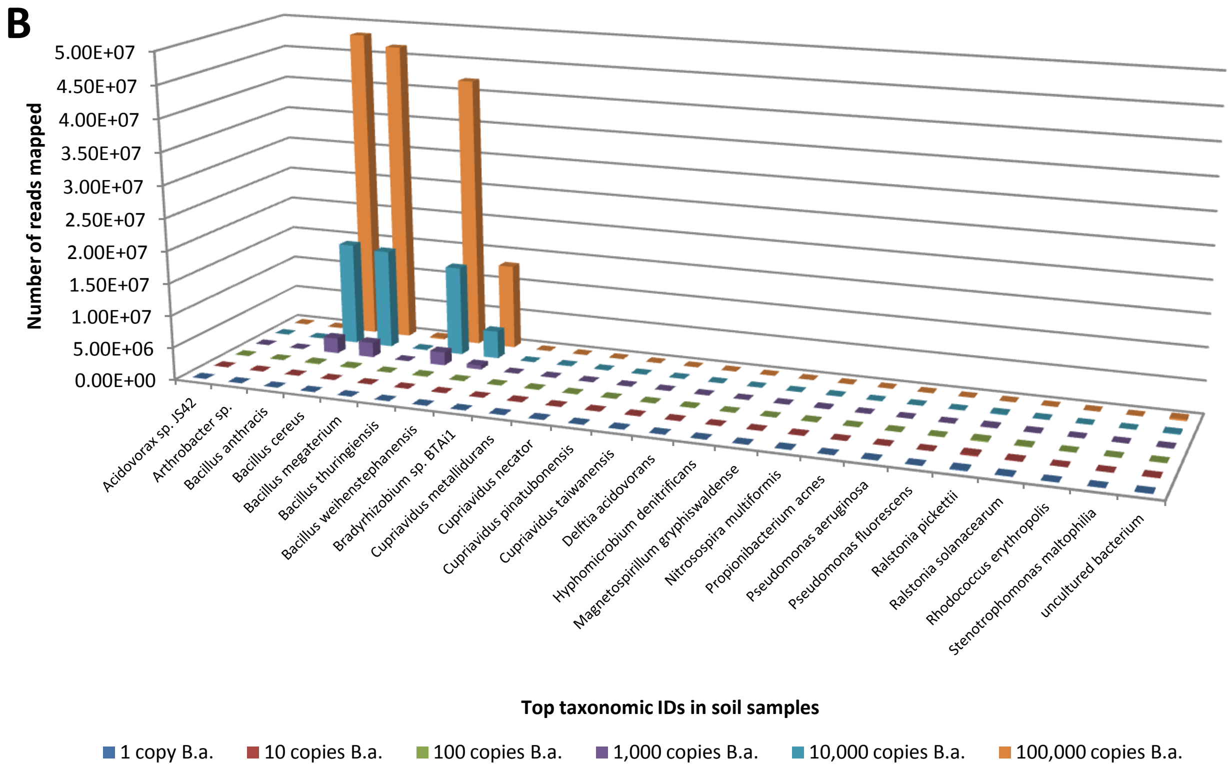
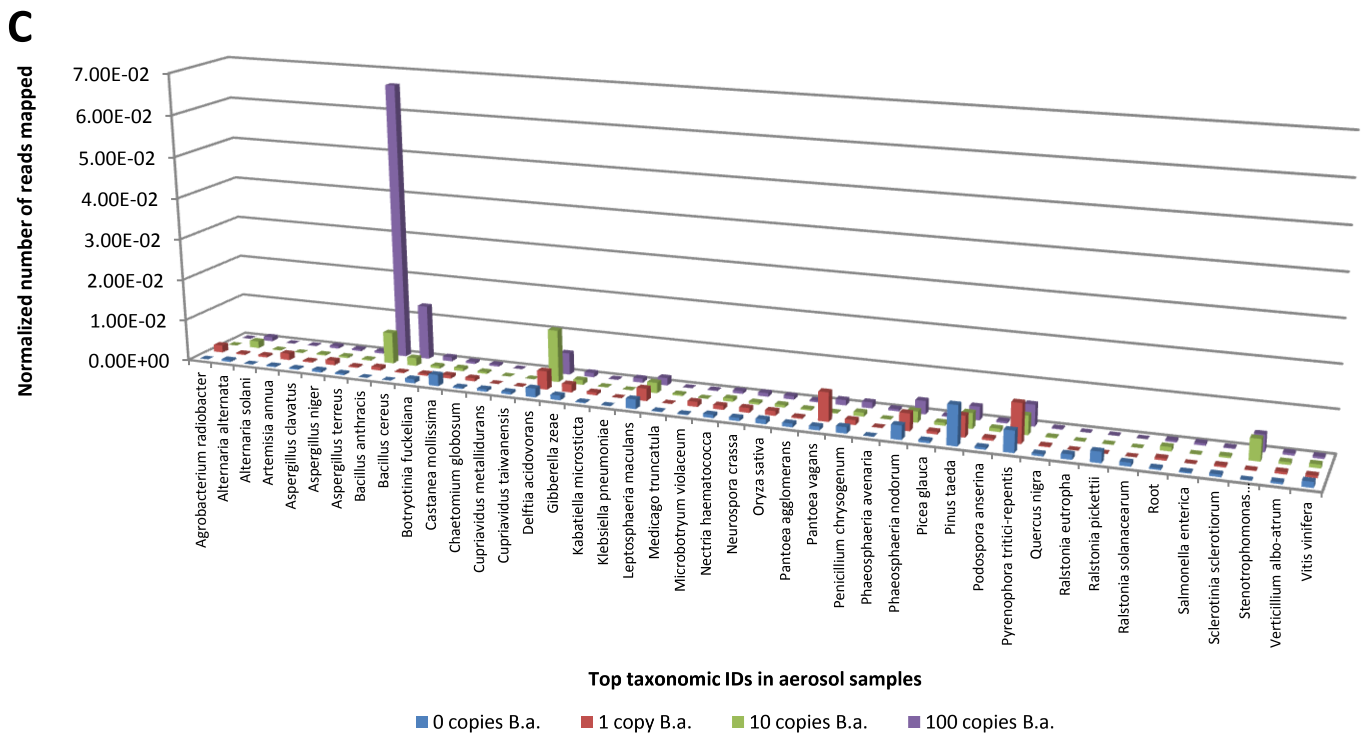

A
Log unique mapped reads
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0
1
10
100
Genomic copies in aerosol sample
■ Bacillus anthracis Ames
■ Bacillus thuringiensis Al hakam
B
Log unique mapped reads
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0
1
10
100
Genomic copies in soil sample
■ Bacillus anthracis Ames
■ Bacillus thuringiensis Al hakam
C
Log unique mapped reads
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0
1
10
100
Genomic copies in aerosol sample
■ Bacillus anthracis Ames
■ Bacillus cereus biovar anthracis CI
D
Log unique mapped reads
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0
1
10
100
Genomic copies in soil sample
■ Bacillus anthracis Ames
■ Bacillus cereus biovar anthracis CI

Figure 5. Mapping of sequencing reads to the full Genbank reference database. We combined in one union set the species from each sample with the highest number of total mapped reads, sorted by number of reads mapping only one taxID. Each species is given with its corresponding number of mapped reads. Shown are the species identified by Illumina sequencing in **A.** aerosol and **B.** soil background samples and the species identified by 454 sequencing in **C.** aerosol and **D.** soil samples.


A
Number of reads mapped
5.00E+07
4.50E+07
4.00E+07
3.50E+07
3.00E+07
2.50E+07
2.00E+07
1.50E+07
1.00E+07
5.00E+06
0.00E+00
Acidovorax sp. JS42
Bacillus anthracis
Bacillus atrophaeus
Bacillus cereus
Bacillus thuringiensis
Bacillus weihenstephanensis
Bradyrhizobium japonicum
Bradyrhizobium sp. BTAi1
Bradyrhizobium sp. ORS278
Cupriavidus metallidurans
Cupriavidus necator
Cupriavidus pinatubonensis
Cupriavidus taiwanensis
Delftia acidovorans
Hyphomicrobium denitrificans
Pantoea vagans
Propionibacterium acnes
Pseudomonas aeruginosa
Ralstonia pickettii
Ralstonia solanacearum
Rhodopseudomonas palustris
Stenotrophomonas maltophilia
Top taxonomic IDs in aerosol samples
1 copy B.a.
10 copies B.a.
100 copies B.a.
1,000 copies B.a.
10,000 copies B.a.
100,000 copies B.a.


B
Number of reads mapped
5.00E+07
4.50E+07
4.00E+07
3.50E+07
3.00E+07
2.50E+07
2.00E+07
1.50E+07
1.00E+07
5.00E+06
0.00E+00
Acidovorax sp. JS42
Arthrobacter sp.
Bacillus anthracis
Bacillus cereus
Bacillus megaterium
Bacillus thuringiensis
Bacillus weihenstephanensis
Bradyrhizobium sp. BTAi1
Cupriavidus metallidurans
Cupriavidus necator
Cupriavidus pinatubonensis
Cupriavidus taiwanensis
Delftia acidovorans
Hyphomicrobium denitrificans
Magnetospirillum gryphiswaldense
Nitrosospira multiformis
Propionibacterium acnes
Pseudomonas aeruginosa
Pseudomonas fluorescens
Ralstonia pickettii
Ralstonia solanacearum
Rhodococcus erythropolis
Stenotrophomonas maltophilia
uncultured bacterium
Top taxonomic IDs in soil samples
1 copy B.a.
10 copies B.a.
100 copies B.a.
1,000 copies B.a.
10,000 copies B.a.
100,000 copies B.a.


C
Normalized number of reads mapped
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
Agrobacterium radiobacter
Alternaria alternata
Alternaria solani
Artemisia annua
Aspergillus clavatus
Aspergillus niger
Aspergillus terreus
Bacillus anthracis
Bacillus cereus
Botryotinia fuckeliana
Castanea mollissima
Chaetomium globosum
Cupriavidus metallidurans
Cupriavidus taiwanensis
Delftia acidovorans
Gibberella zeae
Kabatiella microsticta
Klebsiella pneumoniae
Leptosphaeria maculans
Medicago truncatula
Microbotryum violaceum
Nectria haematococca
Neurospora crassa
Oryza sativa
Pantoea agglomerans
Pantoea vagans
Penicillium chrysogenum
Phaeosphaeria avenaria
Phaeosphaeria nodorum
Picea glauca
Pinus taeda
Podospora anserina
Pyrenophora tritici-repentis
Quercus nigra
Ralstonia eutropha
Ralstonia pickettii
Ralstonia solanacearum
Root
Salmonella enterica
Sclerotinia sclerotiorum
Stenotrophomonas...
Verticillium albo-atrum
Vitis vinifera
Top taxonomic IDs in aerosol samples
🟦 0 copies B.a.
🟥 1 copy B.a.
🟩 10 copies B.a.
🟪 100 copies B.a.

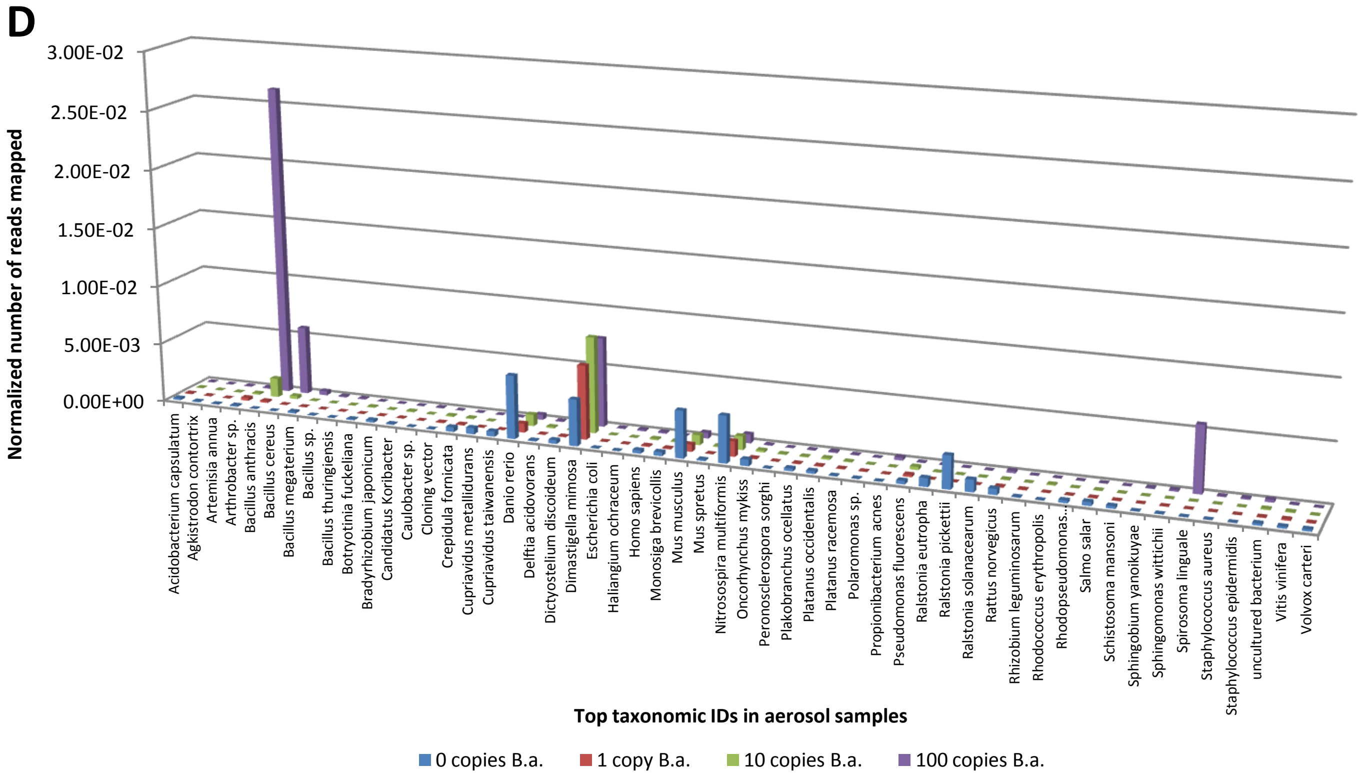

D
Normalized number of reads mapped
3.00E-02
2.50E-02
2.00E-02
1.50E-02
1.00E-02
5.00E-03
0.00E+00
Acidobacterium capsulatum
Agkistrodon contortrix
Artemisia annua
Arthrobacter sp.
Bacillus anthracis
Bacillus cereus
Bacillus megaterium
Bacillus sp.
Bacillus thuringiensis
Botryotinia fuckeliana
Bradyrhizobium japonicum
Candidatus Koribacter
Caulobacter sp.
Cloning vector
Crepidula fornicata
Cupriavidus metallidurans
Cupriavidus taiwanensis
Danio rerio
Delftia acidovorans
Dictyostelium discoideum
Dimastigella mimosa
Escherichia coli
Haliangium ochraceum
Homo sapiens
Monosiga brevicollis
Mus musculus
Mus spretus
Nitrosospira multiformis
Oncorhynchus mykiss
Peronosclerospora sorghi
Plakobranchus ocellatus
Platanus occidentalis
Platanus racemosa
Polaromonas sp.
Propionibacterium acnes
Pseudomonas fluorescens
Ralstonia eutropha
Ralstonia pickettii
Ralstonia solanacearum
Rattus norvegicus
Rhizobium leguminosarum
Rhodococcus erythropolis
Rhodopseudomonas...
Salmo salar
Schistosoma mansoni
Sphingobium yanoikuyae
Sphingomonas wittichii
Spirosoma linguale
Staphylococcus aureus
Staphylococcus epidermidis
uncultured bacterium
Vitis vinifera
Volvox carteri
Top taxonomic IDs in aerosol samples
🟦 0 copies B.a.
🟥 1 copy B.a.
🟩 10 copies B.a.
🟪 100 copies B.a.

Table 3. Detection of *B. anthracis* in environmental background by census microarray. We processed *B. anthracis*-spiked environmental samples using a census array designed to provide broad microbial detection.

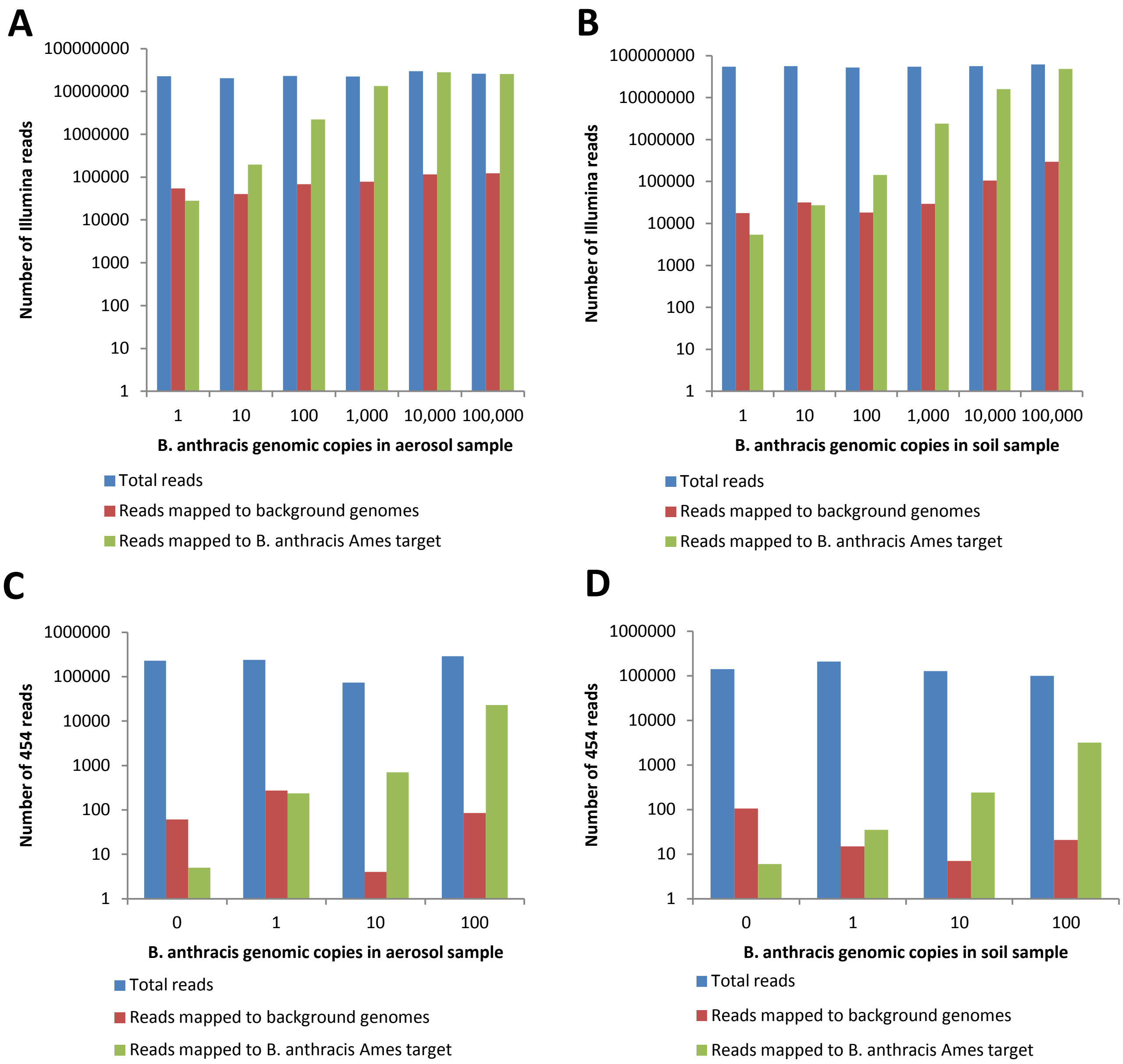
* Detection of *B. anthracis* by only one array replicate

***B. anthracis genome* copy #**	100,000	10,000	1,000	100	10	1
	Aerosol background					
Amount aerosol filter DNA	100 pg	100 pg	100 pg	100 pg	100 pg	100 pg
Array top hit	*B. anthracis*	*B. anthracis*	*B. anthracis*	*B. anthracis*	*B. anthracis**	N/D
	Soil background					
Amount soil DNA	1 ng	1 ng	1 ng	1 ng	1 ng	1 ng
Array top hit	*B. anthracis*	*B. anthracis*	*B. anthracis*	*B. anthracis**	N/D	N/D

Table 4. Summary of detection limits for *B. anthracis* in environmental samples.

	Limit of detection (genome copies/ng environmental DNA)	
Assay	**Aerosol background**	**Soil background**
Illumina sequencing (BWA)	100	100
454 sequencing (gsMapper)	100	10
Microarray	1000	1000

Supplementary Figure S1. Absolute numbers of sequencing reads mapping to specified reference genomes. We spiked increasing genome copy numbers into samples at 10-fold concentration intervals. Following sequencing of *B. anthracis*-spiked samples via Illumina and 454 platforms, we mapped resultant reads to either a target set (*B. anthracis*) or a background set. Shown are total number of sequencing reads from each sample with numbers of reads mapped to reference groups (logarithmic scale) for **A.** Illumina reads from aerosol background, **B.** Illumina reads from soil background, **C.** 454 reads from aerosol background, and **D.** 454 reads from soil background.


A
Number of Illumina reads
100000000
10000000
1000000
100000
10000
1000
100
10
1
1
10
100
1,000
10,000
100,000
B. anthracis genomic copies in aerosol sample
🟦 Total reads
🟥 Reads mapped to background genomes
🟩 Reads mapped to B. anthracis Ames target
B
Number of Illumina reads
100000000
10000000
1000000
100000
10000
1000
100
10
1
1
10
100
1,000
10,000
100,000
B. anthracis genomic copies in soil sample
🟦 Total reads
🟥 Reads mapped to background genomes
🟩 Reads mapped to B. anthracis Ames target
C
Number of 454 reads
1000000
100000
10000
1000
100
10
1
0
1
10
100
B. anthracis genomic copies in aerosol sample
🟦 Total reads
🟥 Reads mapped to background genomes
🟩 Reads mapped to B. anthracis Ames target
D
Number of 454 reads
1000000
100000
10000
1000
100
10
1
0
1
10
100
B. anthracis genomic copies in soil sample
🟦 Total reads
🟥 Reads mapped to background genomes
🟩 Reads mapped to B. anthracis Ames target

Supplementary Table S1. Top 15 species detected in aerosol samples spiked with *B. anthracis*, sorted by number of Illumina reads mapped by Bowtie to only one bacterial species (*B. anthracis* hits in bold).

1 copy B.a	**10 copies B.a**	**100 copies B.a.**	**1,000 copies B.a.**	**10,000 copies B.a**	**100,000 copies B.a.**
Ralstonia pickettii	Ralstonia pickettii	Ralstonia pickettii	**Bacillus anthracis**	**Bacillus anthracis**	**Bacillus anthracis**
Cupriavidus metallidurans	Cupriavidus metallidurans	**Bacillus anthracis**	Ralstonia pickettii	Ralstonia pickettii	Ralstonia pickettii
Ralstonia solanacearum	Ralstonia solanacearum	Cupriavidus metallidurans	Cupriavidus metallidurans	Cupriavidus metallidurans	Bacillus cereus
Bradyrhizobium sp. BTAi1	Delftia acidovorans	Ralstonia solanacearum	Bradyrhizobium sp. BTAi1	Bacillus cereus	Delftia acidovorans
Bradyrhizobium japonicum	Cupriavidus necator	Bradyrhizobium sp. BTAi1	Ralstonia solanacearum	Ralstonia solanacearum	Cupriavidus metallidurans
Delftia acidovorans	Bradyrhizobium sp. BTAi1	Bradyrhizobium japonicum	Bradyrhizobium japonicum	Delftia acidovorans	Bacillus thuringiensis
Rhodopseudomonas palustris	Bradyrhizobium japonicum	Delftia acidovorans	Delftia acidovorans	Bradyrhizobium sp. BTAi1	Propionibacterium acnes
Cupriavidus necator	Cupriavidus taiwanensis	Rhodopseudomonas palustris	Rhodopseudomonas palustris	Bradyrhizobium japonicum	Ralstonia solanacearum
Cupriavidus taiwanensis	**Bacillus anthracis**	Cupriavidus necator	Hyphomicrobium denitrificans	Hyphomicrobium denitrificans	Bradyrhizobium sp. BTAi1
Cupriavidus pinatubonensis	Cupriavidus pinatubonensis	Cupriavidus taiwanensis	Cupriavidus necator	Rhodopseudomonas palustris	Bacillus atrophaeus
Hyphomicrobium denitrificans	Rhodopseudomonas palustris	Cupriavidus pinatubonensis	Cupriavidus taiwanensis	Bacillus thuringiensis	Bacillus weihenstephanensis
Bradyrhizobium sp. ORS278	Pseudomonas aeruginosa	Bradyrhizobium sp. ORS278	Bradyrhizobium sp. ORS278	Cupriavidus taiwanensis	Bradyrhizobium japonicum
Pantoea vagans	Hyphomicrobium denitrificans	Acidovorax sp. JS42	Cupriavidus pinatubonensis	Cupriavidus necator	Rhodopseudomonas palustris
Pseudomonas aeruginosa	Stenotrophomonas maltophilia	Hyphomicrobium denitrificans	Stenotrophomonas maltophilia	Cupriavidus pinatubonensis	Cupriavidus taiwanensis
Stenotrophomonas maltophilia	Bradyrhizobium sp. ORS278	Pantoea vagans	Pseudomonas aeruginosa	Bacillus weihenstephanensis	Cupriavidus necator

Supplementary Table S2. Top 15 species detected in soil samples spiked with *B. anthracis*, sorted by number of Illumina reads mapped by Bowtie to only one bacterial species (*B. anthracis* hits in bold).

1 copy B.a	**10 copies B.a**	**100 copies B.a.**	**1,000 copies B.a.**	**10,000 copies B.a**	**100,000 copies B.a.**
Ralstonia pickettii	Ralstonia pickettii	Ralstonia pickettii	Ralstonia pickettii	**Bacillus anthracis**	**Bacillus anthracis**
Nitrosospira multiformis	Nitrosospira multiformis	Nitrosospira multiformis	**Bacillus anthracis**	Ralstonia pickettii	Ralstonia pickettii
Cupriavidus metallidurans	Cupriavidus metallidurans	Cupriavidus metallidurans	Nitrosospira multiformis	Nitrosospira multiformis	Nitrosospira multiformis
Ralstonia solanacearum	Ralstonia solanacearum	Ralstonia solanacearum	Cupriavidus metallidurans	Cupriavidus metallidurans	Cupriavidus metallidurans
Delftia acidovorans	Delftia acidovorans	Delftia acidovorans	Ralstonia solanacearum	Ralstonia solanacearum	Bacillus cereus
Cupriavidus necator	Cupriavidus necator	Cupriavidus necator	Delftia acidovorans	Cupriavidus necator	Ralstonia solanacearum
Cupriavidus taiwanensis	Cupriavidus taiwanensis	**Bacillus anthracis**	Cupriavidus necator	Delftia acidovorans	Delftia acidovorans
Cupriavidus pinatubonensis	Hyphomicrobium denitrificans	Cupriavidus taiwanensis	Rhodococcus erythropolis	Cupriavidus taiwanensis	Cupriavidus necator
Stenotrophomonas maltophilia	Cupriavidus pinatubonensis	Bacillus megaterium	Cupriavidus taiwanensis	Propionibacterium acnes	Cupriavidus taiwanensis
Hyphomicrobium denitrificans	Arthrobacter sp.	Rhodococcus erythropolis	Hyphomicrobium denitrificans	Cupriavidus pinatubonensis	Bacillus thuringiensis
uncultured bacterium	Bacillus megaterium	Cupriavidus pinatubonensis	Cupriavidus pinatubonensis	Hyphomicrobium denitrificans	Cupriavidus pinatubonensis
Pseudomonas aeruginosa	uncultured bacterium	Hyphomicrobium denitrificans	uncultured bacterium	Pseudomonas fluorescens	Hyphomicrobium denitrificans
Magnetospirillum gryphiswaldense	Pseudomonas aeruginosa	uncultured bacterium	Bradyrhizobium sp. BTAi1	uncultured bacterium	Arthrobacter sp.
Pseudomonas fluorescens	Stenotrophomonas maltophilia	Acidovorax sp. JS42	Bacillus megaterium	Bacillus cereus	Bacillus weihenstephanensis
Acidovorax sp. JS42	Bradyrhizobium sp. BTAi1	Pseudomonas aeruginosa	Magnetospirillum gryphiswaldense	Acidovorax sp. JS42	Stenotrophomonas maltophilia

Supplementary Table S3. Summary of microbial sequences represented on the census array.

Number of Targets	***Viral***	***Bacterial***
Families	80	274
Groups without family classification	48	65
Species with complete genome, plasmid, or segment data	2530	1290
Species with sequence data, including sequence fragments	5719	14765
Sequences classified as to Family	171264	728467
Sequences unclassified as to Family	6996	56251
Complete genomes, segments, or plasmids	55803	4122

REFERENCES

1. Driks A. The Bacillus anthracis spore. Mol Aspects Med 2009;30:368-73

2. Read TD, Peterson SN, Tourasse N, et al. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. Nature 2003;423:81-6

3. Young JA, Collier RJ. Anthrax toxin: receptor binding, internalization, pore formation, and translocation. Annu Rev Biochem 2007;76:243-65

4. Jernigan DB, Raghunathan PL, Bell BP, et al. Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. Emerg Infect Dis 2002;8:1019-28

5. Peters CJ, Hartley DM. Anthrax inhalation and lethal human infection. Lancet 2002;359:710-1

6. Song Y, Yang R, Guo Z, Zhang M, Wang X and Zhou F. Distinctness of spore and vegetative cellular fatty acid profiles of some aerobic endospore-forming bacilli. J Microbiol Methods 2000;39:225-41

7. Keys CJ, Dare DJ, Sutton H, et al. Compilation of a MALDI-TOF mass spectral database for the rapid screening and characterisation of bacteria implicated in human infectious diseases. Infect Genet Evol 2004;4:221-42

8. Irenge LM, Gala JL. Rapid detection methods for Bacillus anthracis in environmental samples: a review. Appl Microbiol Biotechnol 2012

9. Quinn CP, Semenova VA, Elie CM, et al. Specific, sensitive, and quantitative enzyme-linked immunosorbent assay for human immunoglobulin G antibodies to anthrax toxin protective antigen. Emerg Infect Dis 2002;8:1103-10

10. Kumar S, Tuteja U. Detection of virulence-associated genes in clinical isolates of bacillus anthracis by multiplex PCR and DNA probes. J Microbiol Biotechnol 2009;19:1475-81

11. Rao SS, Mohan KV and Atreya CD. Detection technologies for Bacillus anthracis: prospects and challenges. J Microbiol Methods 2010;82:1-10

12. Gardner SN, Jaing CJ, McLoughlin KS and Slezak TR. A microbial detection array (MDA) for viral and bacterial detection. BMC Genomics 2010;11:668

13. Herzog AB, McLennan SD, Pandey AK, et al. Implications of limits of detection of various methods for Bacillus anthracis in computing risks to human health. Appl Environ Microbiol 2009;75:6331-9

14. Cummings CA, Bormann Chung CA, Fang R, et al. Accurate, rapid and high-throughput detection of strain-specific polymorphisms in Bacillus anthracis and Yersinia pestis by next-generation sequencing. Investig Genet 2010;1:5

15. Rasko DA, Worsham PL, Abshire TG, et al. Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. Proc Natl Acad Sci U S A 2011;108:5027-32

16. Wright AM, Beres SB, Consamus EN, et al. Rapidly progressive, fatal, inhalation anthrax-like infection in a human: case report, pathogen genome sequencing, pathology, and coordinated response. Arch Pathol Lab Med 2011;135:1447-59

17. Chen PE, Willner KM, Butani A, et al. Rapid identification of genetic modifications in Bacillus anthracis using whole genome draft sequences generated by 454 pyrosequencing. PLoS One;5:e12397

18. Helgason E, Okstad OA, Caugant DA, et al. Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis--one species on the basis of genetic evidence. Appl Environ Microbiol 2000;66:2627-30

19. Hoffmaster AR, Hill KK, Gee JE, et al. Characterization of Bacillus cereus isolates associated with fatal pneumonias: strains are closely related to Bacillus anthracis and harbor B. anthracis virulence genes. J Clin Microbiol 2006;44:3352-60

20. Kolsto AB, Tourasse NJ and Okstad OA. What sets Bacillus anthracis apart from other Bacillus species? Annu Rev Microbiol 2009;63:451-76

21. Klee SR, Ozel M, Appel B, et al. Characterization of Bacillus anthracis-like bacteria isolated from wild great apes from Cote d'Ivoire and Cameroon. J Bacteriol 2006;188:5333-44

22. Beyer W, Pocivalsek S and Bohm R. Polymerase chain reaction-ELISA to detect Bacillus anthracis from soil samples-limitations of present published primers. J Appl Microbiol 1999;87:229-36

23. Tims TB, Lim DV. Rapid detection of Bacillus anthracis spores directly from powders with an evanescent wave fiber-optic biosensor. J Microbiol Methods 2004;59:127-30

24. Estill CF, Baron PA, Beard JK, et al. Comparison of air sampling methods for aerosolized spores of B. anthracis Sterne. J Occup Environ Hyg 2011;8:179-86

25. Stratis-Cullum DN, Griffin GD, Mobley J, Vass AA and Vo-Dinh T. A miniature biochip system for detection of aerosolized Bacillus globigii spores. Anal Chem 2003;75:275-80

26. Giegerich R, Kurtz S and Stoye J. Efficient implementation of lazy suffix trees. Software-Practice and Experience 2003;33:1035-1049

27. Jaing C, Gardner S, McLoughlin K, et al. A functional gene array for detection of bacterial virulence elements. PLoS One 2008;3:e2163

28. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, eds. Bioinformatics Methods and Protocols: Methods in Molecular Biology Totowa, NJ: Humana Press, 2000:365-386

29. Markham NR, Zuker M. DNAMelt web server for nucleic acid melting prediction. Nucleic Acids Res. 2005;33:W577-W581